Los Alamos
NATIONAL LABORATORY
EST 1943

New AI tool makes vast data streams intelligible and explainable

March 11, 2021

LOS ALAMOS, N.M., March 11, 2021—Making sense of vast streams of big data is getting easier, thanks to an artificial-intelligence tool developed at Los Alamos National Laboratory. SmartTensors sifts through millions of millions of bytes of diverse data to find the hidden features that matter, with significant implications from health care to national security, climate modeling to text mining, and many other fields.

“SmartTensors analyzes terabytes of diverse data to find the hidden patterns and features that make the data understandable and reveal its underlying processes or causes,” said Boian Alexandrov, a scientist at Los Alamos National Laboratory, AI expert, and principal investigator on the project. “What our AI software does best is extracting the latent variables, or features, from the data that describe the whole data and mechanisms buried in it without any preconceived hypothesis.”

SmartTensors also can identify the optimal number of features needed to make sense of enormous, multidimensional datasets.

“Finding the optimal number of features is a way to reduce the dimensions in the data while being sure you’re not leaving out significant parts that lead to understanding the underlying processes shaping the whole the dataset,” said Velimir (“Monty”) Vesselinov, an expert in machine learning, data analytics and model diagnostics at Los Alamos and also a principal investigator.

Tensors make big data manageable

The nonnegativity of the latent features and determining their optimal number reduce a vast data set to a scale that’s manageable for computers to process and subject-matter-experts to analyze. The extracted features are explainable and understandable.

Features are discrete, intelligible chunks of data. For instance, in a database of human faces, key features are noses, eyes, eyebrows, ears, mouths, and chins. SmartTensors can be pointed at a database of faces and, without human guidance, isolate those features. It also can determine how many of those features—the optimal number—are required to do the job accurately and reliably. For instance, maybe eyebrows and chins are unnecessary for facial recognition.

In other database examples, latent features may represent climate processes, watershed mechanisms, hidden geothermal resources, carbon sequestration processes, chemical reactions, protein structures, pharmaceutical molecules, cancerous mutations in human genomes, and so on. In very large datasets—measured in billions of millions of bytes—these features are frequently unknown and invisible to direct observation, obscured by a torrent of less-useful information and noise presented in the data.

SmartTensors works with the notion of a tensor, or multidimensional data cube. Each axis defining the cube represents a different dimension of the data. So, in a business example, information about customers might be on the X axis, information about annual sales on the Y axis, and information about manufacturing on the Z axis. As more of these features are added, the cube becomes more complex—with more dimensions—than the simple 3D cube. If you think of the data cube as being made up of many small, stacked cubes, each one represents information about some or all of the features, or dimensions of the data.

Awash in data

Our world is awash in a seemingly bottomless ocean of data pouring in from sources ranging from satellites and MRI scans to massive computer simulations and seismic-sensor networks, from electronic surveillance to smart phones, from genome sequencing of SARS-Cov-2 to COVID-19 test results, from social networks to vast number of texts. Making sense of this ever-increasing racket is vital to national security, economic stability, individual health, and practically every branch of science.

These vast data sets are formed exclusively by observable quantities, by definition —think of eyes, noses, and ears. But in big-data analytics, it is difficult to directly link these observables to the underlying processes that generate the data. These processes or hidden features remain latent—they're not directly observable and are confusingly mixed with each other, with unimportant features, and with noise.

The problem is often likened to extracting the individual voices at a noisy cocktail party, with a set of microphones recording the chatter. Isolating conversation or conversations while individuals are walking and talking is a classic signal-processing blind-separation problem. The number of latent features here is the number of individual voices and their characteristics, which might include the pitch and tone of each person's voice, for instance. Once that's determined, it's easier to follow a conversational thread or a person.

In big-data problems, extracting the latent features reveals valuable information about previously unknown causes and mechanisms hidden in the data. Extracting features also uncovers a simplified latent structure that can be used to represent, compress and interpret the entire dataset.

Technical talk

SmartTensors is an unsupervised AI platform based on non-negative tensor factorization and tensor networks. The software uniquely estimates the latent dimension of large (1 terabytes or more), dense and sparse, and diverse datasets. SmartTensors analyzes their structure and identifies hidden physical processes. Running on distributed CPUs, GPUs, and tensor processing units (TPUs), the software was developed in Python, Julia, and C++ and uses open-source and custom libraries. The software works on different platforms from supercomputers like Summit (Oak Ridge National Laboratory) and Sierra (Lawrence Livermore National Laboratory) to usual desktops and even on Quantum Annealers (D-wave, Los Alamos).

The method was applied to a set of applications in various fields, including:

- Medicine: discovering mutational signatures in cancer genomics, biological pathways, and protein structures.
- Economy: performing macroeconomic analyses.
- Chemistry: studying phase separation in complex liquids of co-polymers, cell-membrane rafts, etc.
- Climate: detecting micro-climate biota patterns.
- Material Science: creating combinatorial material X-ray libraries.
- Text Mining: modeling topics.
- Agriculture: estimating the role of different artificial substances on the yield.
- Computer security: detecting anomalies and fraud.
- Blind-source separation: detecting Brownian and Anomaly diffusion of unknown number of sources.
- Relational databases: conducting Boolean factorization analysis of categorical patterns.
- Data Compression: compressing big-data, such as scientific computer-generated data.
- Hydrogeology: pinpointing contamination, contaminant sources, subsurface water pressure patterns.
- Hydrology: characterizing watershed dynamics.
- Geology: discovering hidden geothermal resources, characterizing oil and gas production, and optimizing carbon sequestration.
- Pandemic: identifying and ranking socioeconomic factors impacting disease spreading.
- Wildfires: discovering potential “hot” spots that may trigger large wildfires.
- Seismology: detecting induced seismicity, industrial noise.

For more information, visit the SmartTensors website.

The funding: The SmartTensors software was developed under a Los Alamos project titled Tensor Networks: Robust Unsupervised Machine Learning for Big-Data Analytics, which was funded by the Laboratory Directed Research and Development (LDRD) program at Los Alamos National Laboratory.

LA-UR-21-21940

Los Alamos National Laboratory **www.lanl.gov** **(505) 667-7000** **Los Alamos, NM**

Managed by Triad National Security, LLC for the U.S Department of Energy's NNSA